Workflow in a Federated Earth System Grid Enterprise System

Ann Chervenak[1], Ian Foster[2], and Dean N. Williams[3]
[1]USC Information Sciences Institute; [2]Argone National Laboratory; [3]Lawrence Livermore National Laboratory

Summary

The Earth System Grid Center for Enabling Technologies (ESG-CET) was established to provide community access to hundreds of petabytes of simulation data being generated in the climate community within the next five years. The current ESG-CET project will provide a new, federated architecture that provides for a larger number of distributed sites throughout the world, requiring multiple portals and data delivery mechanisms, a variety of means for user access, and reliable mechanisms to handle system and network failures. The ESG-CET is working with the Center for Enabling Distributed Petascale Science (CEDPS) in several areas, including the development of an analysis service framework; data management functionality, including the use of the GridFTP for the transfer of large climate data sets among ESG sites, and the use of lightweight tools for data replication and mirroring; and monitoring and troubleshooting capabilities.

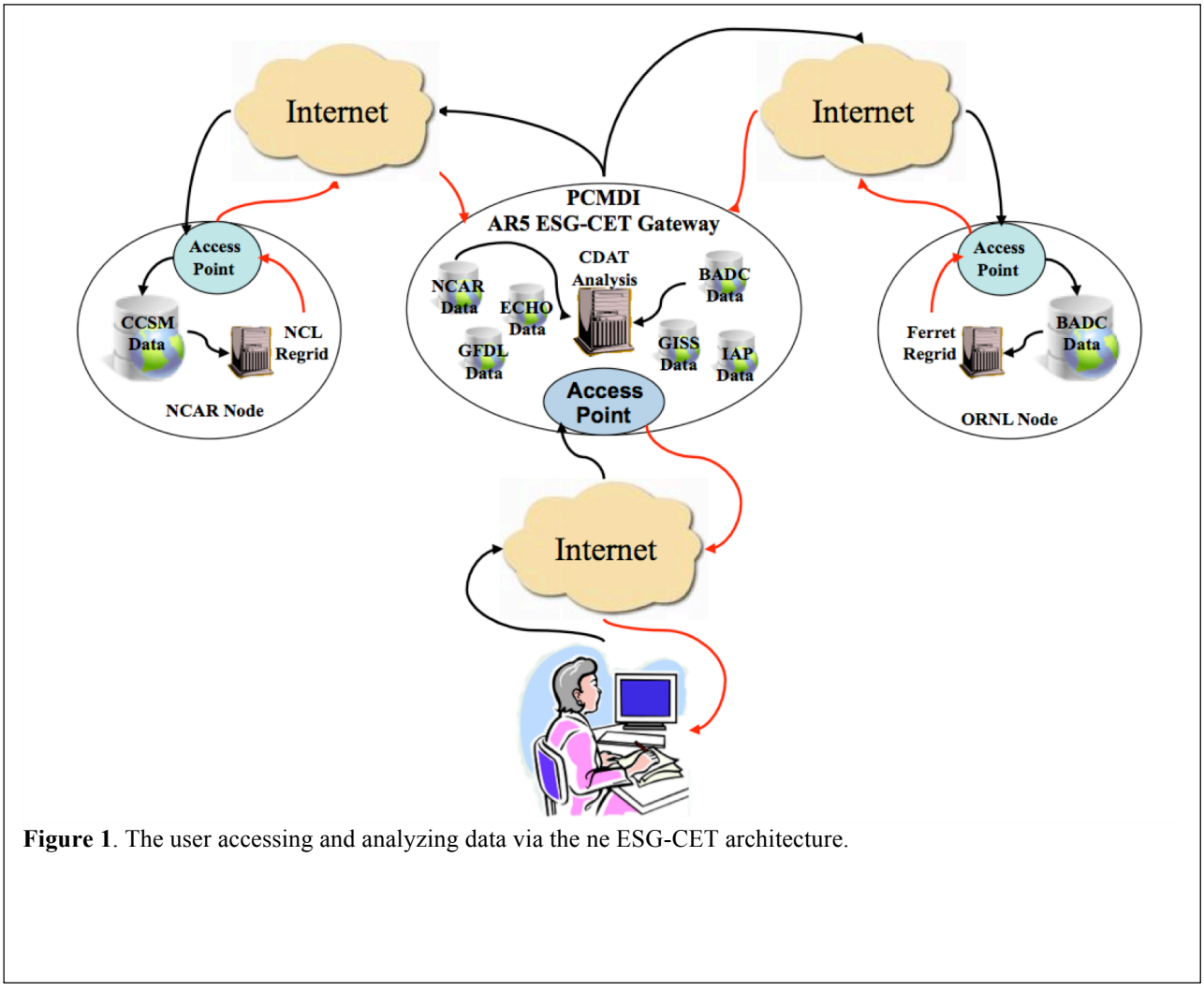

Figure 1. The user accessing and analyzing data via the ne ESG-CET architecture.

Increased computing power and increasingly comprehensive climate models have resulted in a dramatic increase in data output describing the Earth system model. Indeed, over the past decade, model output has increased from megabytes to terabytes, and climatologists expect to generate hundreds of petabytes of simulation data within the next five years. Using this data presents enormous challenges. For example, the data will probably need to be stored in a few archival sites; data users will therefore need assistance in locating data of interest and then in performing data reduction, analysis, and visualization, presumably on the server side to reduce the volume of data sent over the network and stored at the user's location. Users will also need to have confidence in the origins and integrity of the data and the associated metadata.

The Earth System Grid Center for Enabling Technologies (ESG-CET) was established to address such challenges. The ESG-CET builds on the first and highly successful Earth System Grid SciDAC project, which produced an architecture to support a small set of well-known U.S. archive sites operating ESG nodes. For the current project, an entirely new architecture is needed that provides for a larger number of distributed sites throughout the world, requiring multiple portals and data delivery mechanisms, a variety of means for user access, and reliable mechanisms to handle system and network failures. The new framework is based on a three-tiered approach that includes metadata services for search and discovery, data gateways that act as brokers handling data requests to serve specific user communities, and ESG nodes with actual data holdings and metadata accessing services. A distributed testbed based on this new architecture is scheduled to be in place by mid 2009.

The Center works closely with the CEDPS SciDAC project in several areas. First, ESG-CET and CEDPS are working together to develop a basic analysis service framework. The work requires development of techniques to apply existing analysis procedures, written in scripting language systems, to large (many-terabyte) datasets; tools to manage the movement of data between archival storage and disk, since the datasets in storage may be too large to move onto disk in their entirety at one time; and mechanisms to enable users to define and upload their own analysis procedures, without compromising the security or reliability of the server on which those procedures run.

The ESG-CET and CEDPS projects also collaborate in several areas related to data management. ESG-CET uses the GridFTP data transfer service extensively, and thus the project has benefited from the significant performance and functionality enhancements to GridFTP that have been developed under the CEDPS project. In particular, ESG-CET uses the Storage Resource Manager from LBNL to perform wide area bulk data movement of large (terabyte) data sets; SRM in turn calls GridFTP to perform these transfers among ESG sites. The OPeNDAP-G server used in ESG-CET for data access also makes use of GridFTP methods for data transport.

In addition, the ESG-CET and CEDPS projects collaborate on issues related to the replication and mirroring of important climate data sets. Historically, the Earth System Grid has used the Replica Location Service to keep track of the location of files. In the current federated architecture, ESG-CET is working with CEDPS to develop lightweight, efficient tools for mirroring essential data sets among key ESG sites and for keeping track of the locations of these mirrored files using metadata catalogs and standard discovery protocols.

Finally, the ESG-CET and CEDPS projects have common interests related to monitoring and troubleshooting. ESG-CET has used the Globus Monitoring and Discovery System for several years to monitor the status of ESG resources, report failures, facilitate repairs and provide high resource availability. The two projects may in the future share troubleshooting infrastructure being developed by the CEDPS project.

For further information on this subject contact:
Name: Ian T. Foster
Organization: Argonne National Laboratory
Email: Foster@anl.gov
Phone: (630) 252-4619